UNMIX

Theory and Applications

Problem

- Given
  - a data set of compositions of many species for many samples
- With as few assumptions as possible, find
  - the number of sources,
  - the composition of the sources, and
  - the uncertainties.

Physical Basis

- Physical models of source apportionment problems can often written in the same mathematical form as a statistical model, e.g., mass balance and factor analysis:

$$C_{ij} = \sum_{k=1}^{N} a_{jk} S_{ik} + \varepsilon_{ij}\text{, or in matrix terms, } C = SA' + E$$

- C = concentrations, A = source compost-ions,S=source contributions, E=errors, i=1 to n observations, j=1 to m species, k=1 to N sources

The Challenge

- The problem is ill-defined, or not identifiable in the sense that an infinite number of solutions exist that
  - have the same root mean squared error, and
  - satisfy the non-negativity constraints for source compositions and contributions

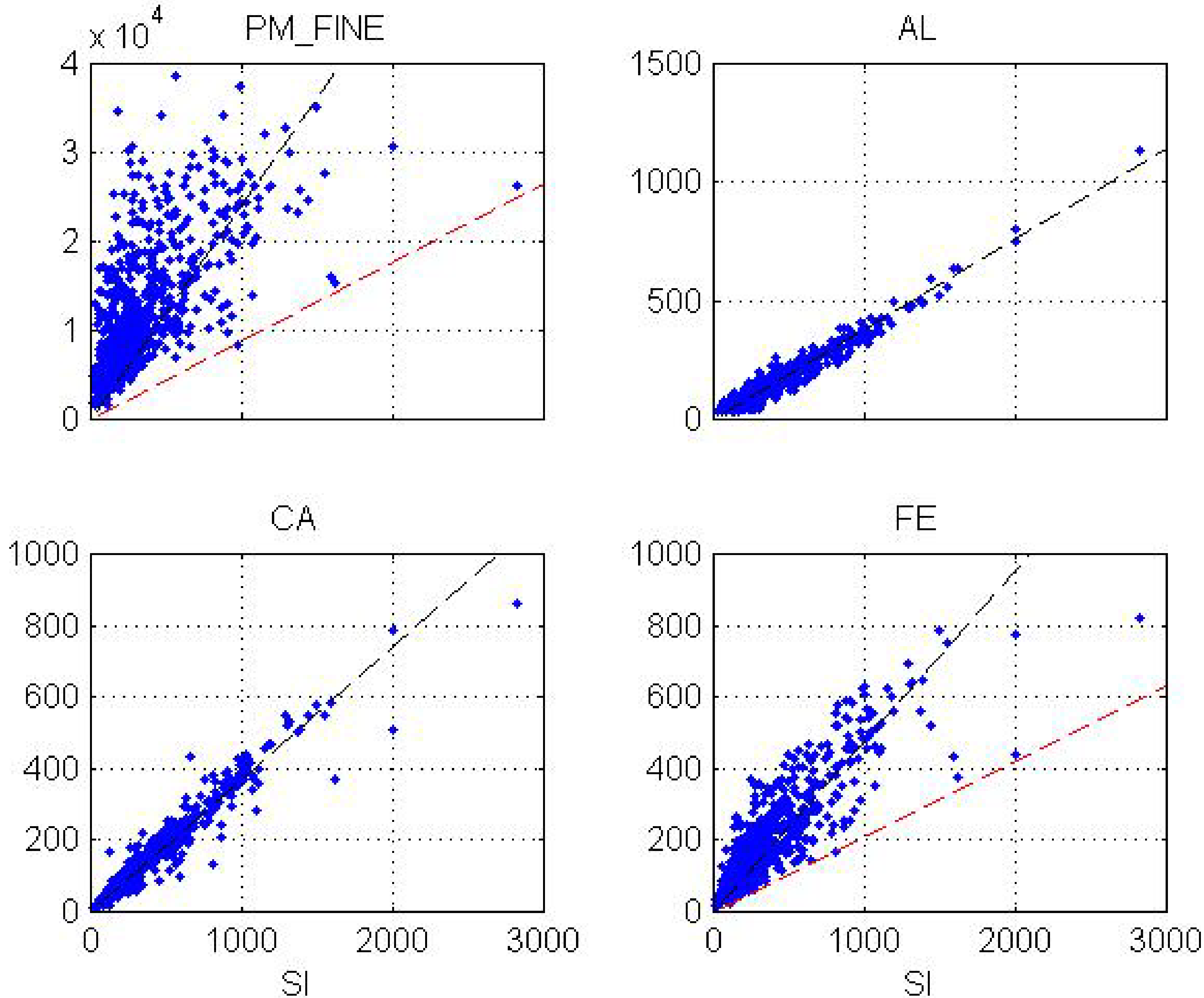
Key Problems in Multivariate Receptor Modeling

- Estimate the number of factors in the data that are present above the noise level
- Find additional constraints for a unique solution.

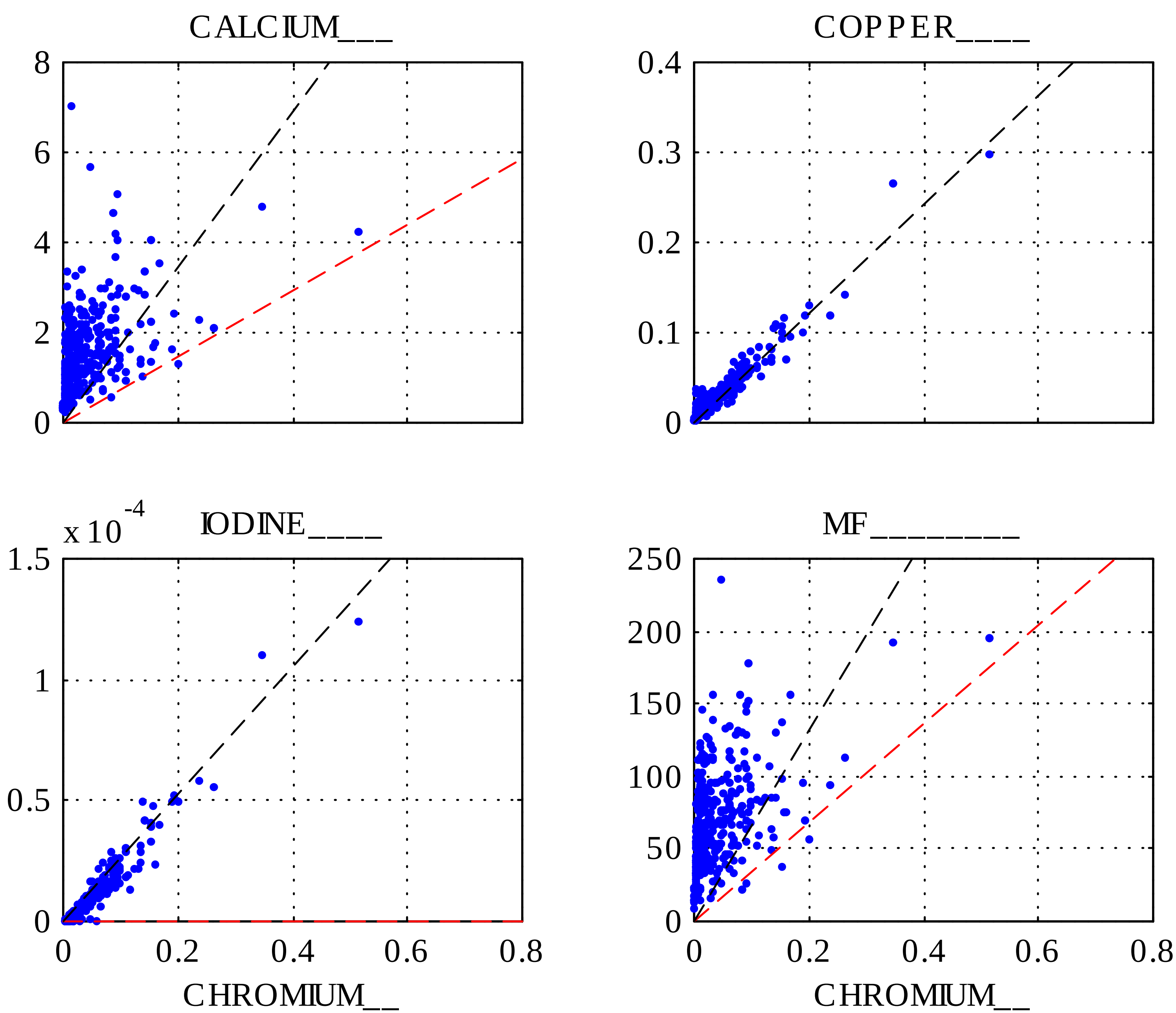
αγεωμετρητοζ μνδειζ εισιτω

Let None Ignorant of Geometry Enter

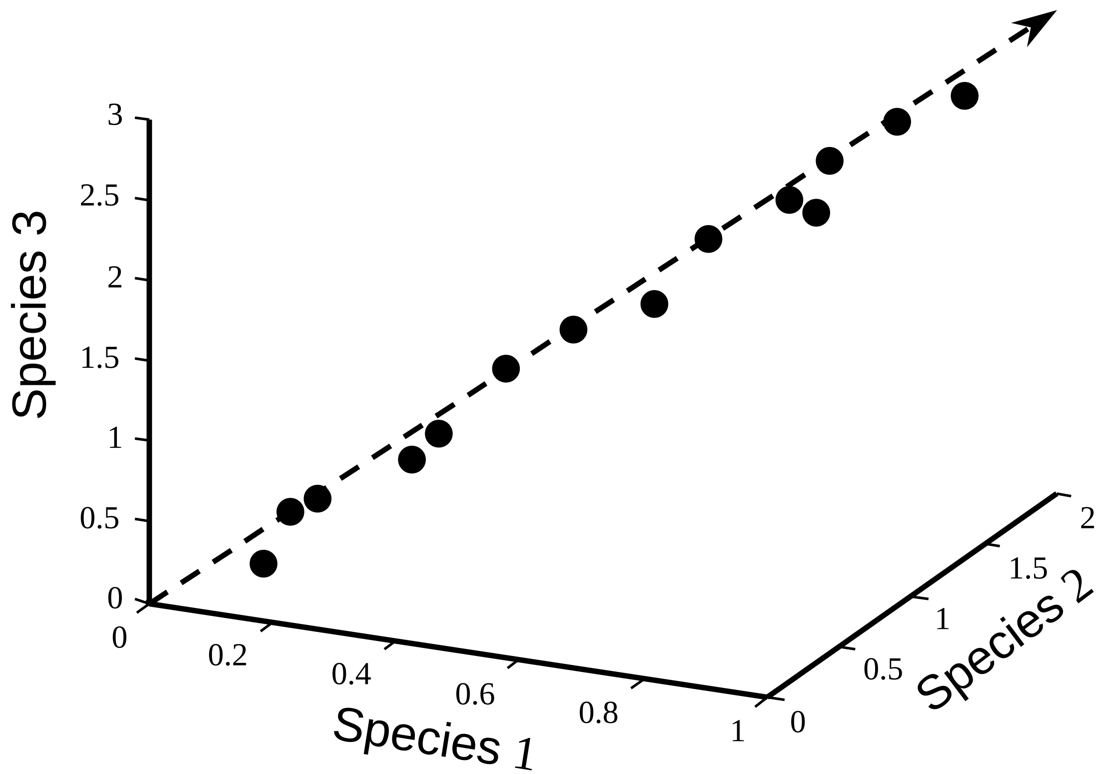
Geometrical Motivation



PM_FINE
x 10^4
4
3
2
1
0
AL
1500
1000
500
0
CA
1000
800
600
400
200
0
0
1000
2000
3000
SI
FE
1000
800
600
400
200
0
0
1000
2000
3000
SI

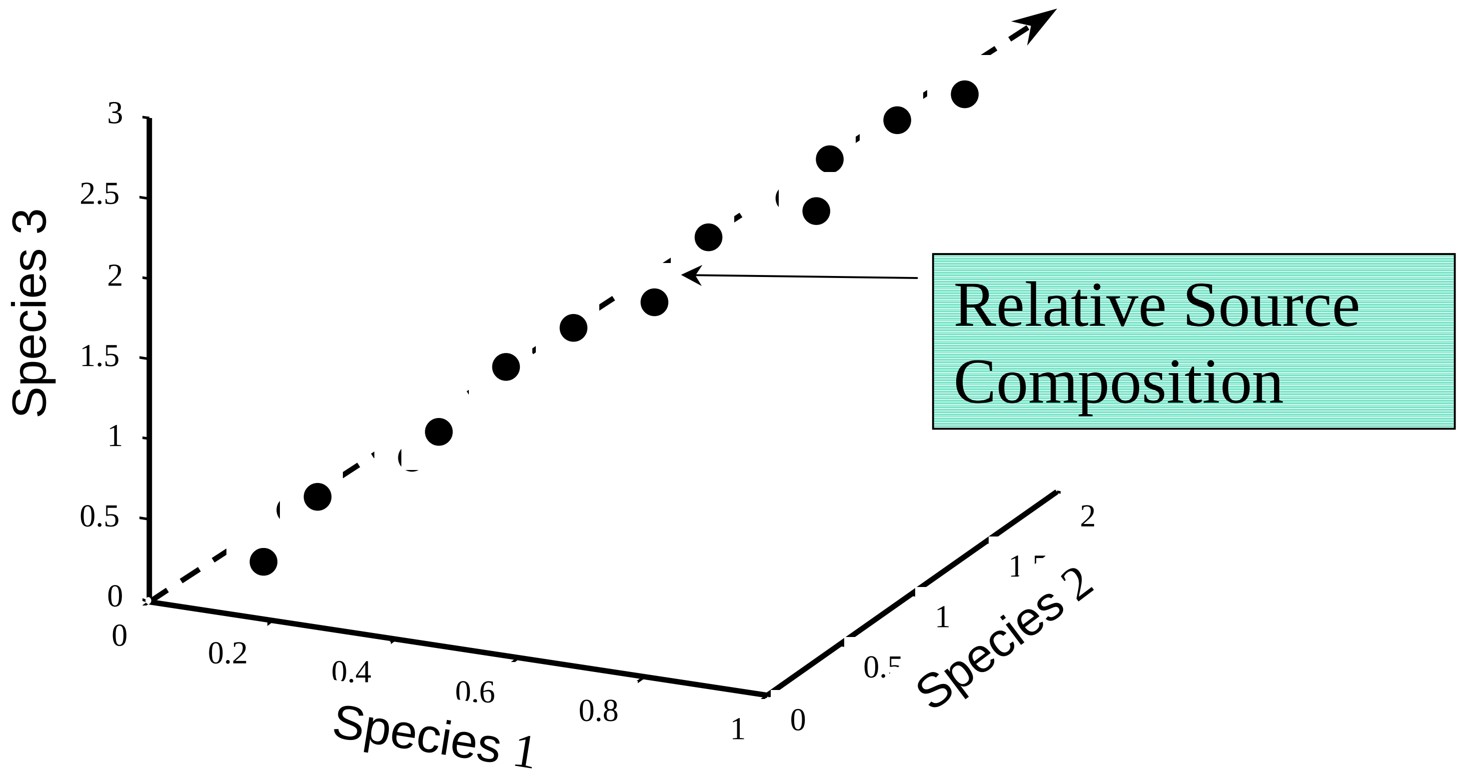

CALCIUM___
COPPER____
IODINE____
MF_________
x 10^-4
8
6
4
2
0
0.4
0.3
0.2
0.1
1.5
1
0.5
250
200
150
100
50
0
0.2
0.4
0.6
0.8
CHROMIUM__
CHROMIUM__

One Source

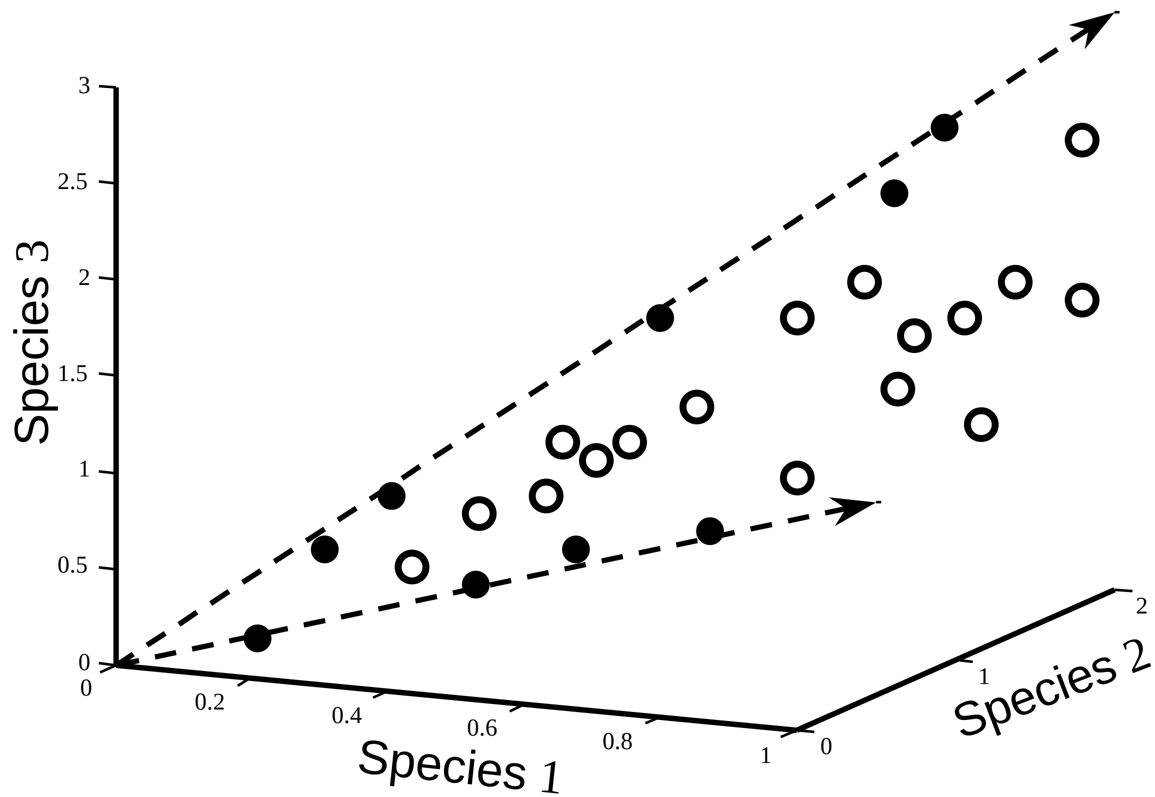


Species 3
3
2.5
2
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1
Species 1
0
0.5
1
1.5
2
Species 2

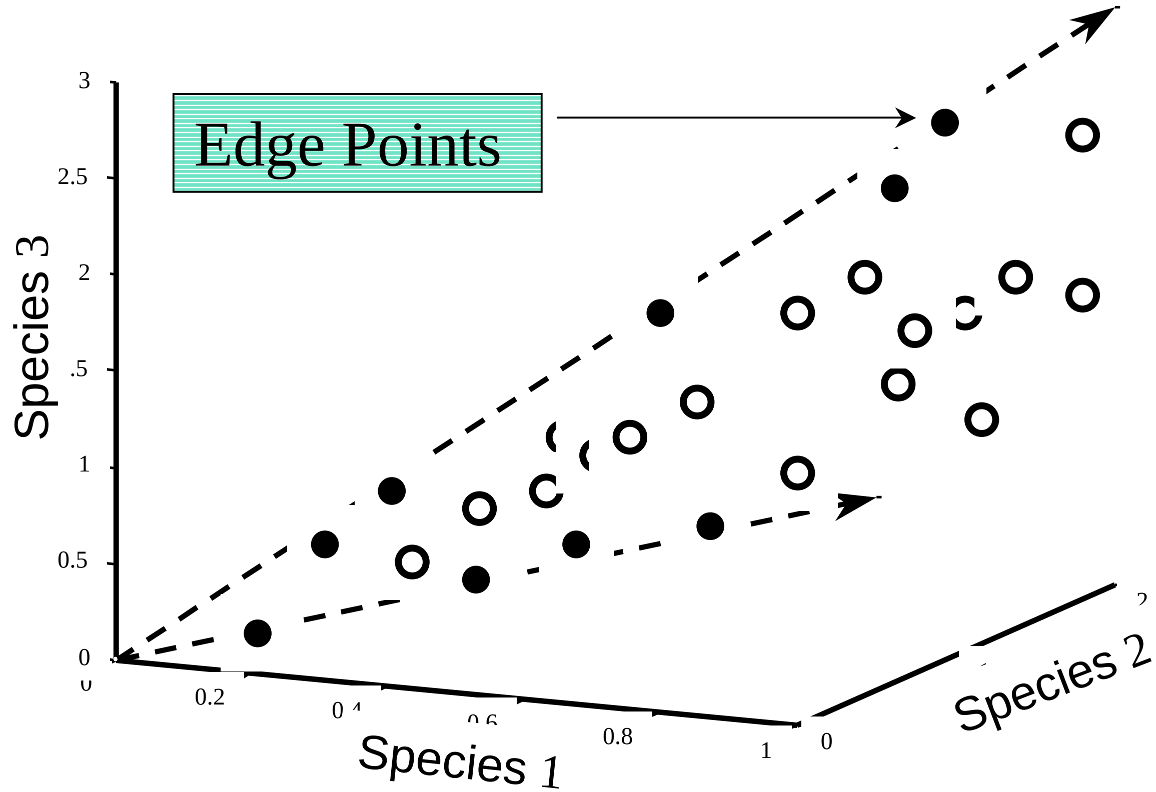
One Source - Line



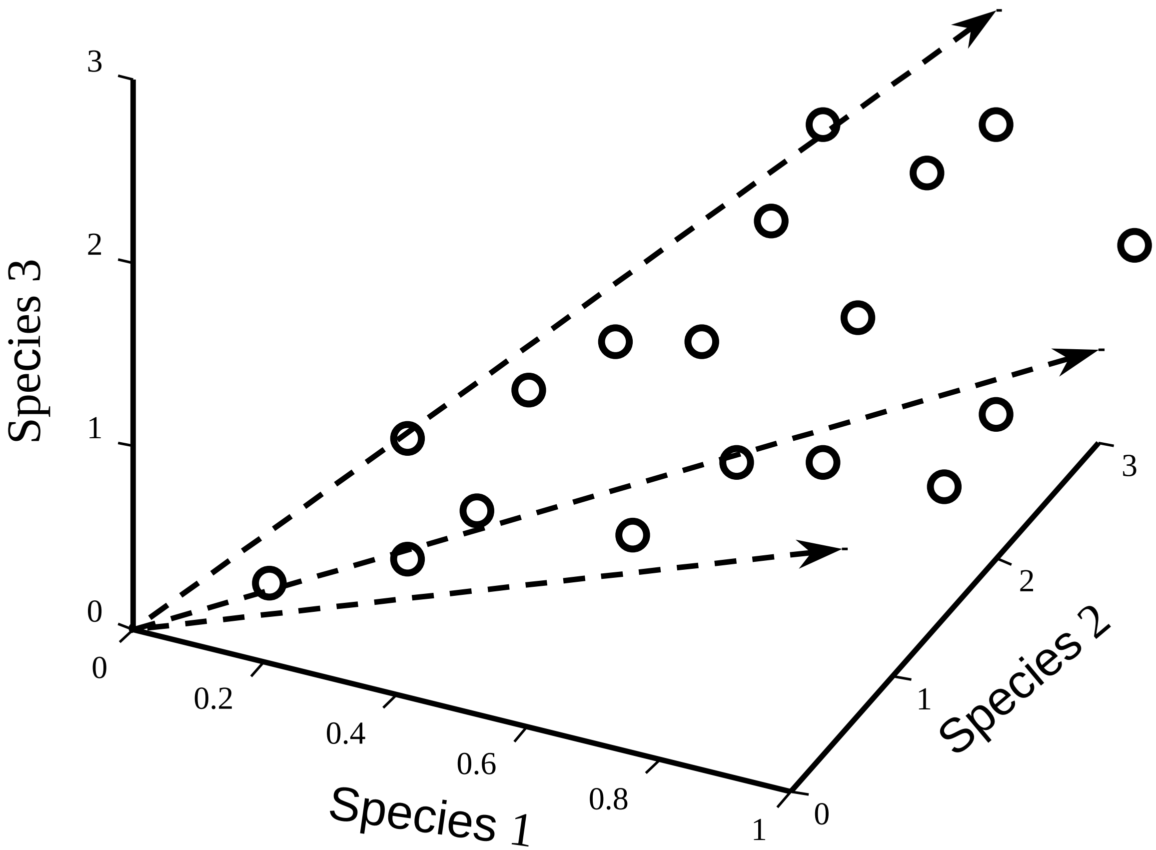
Two Sources



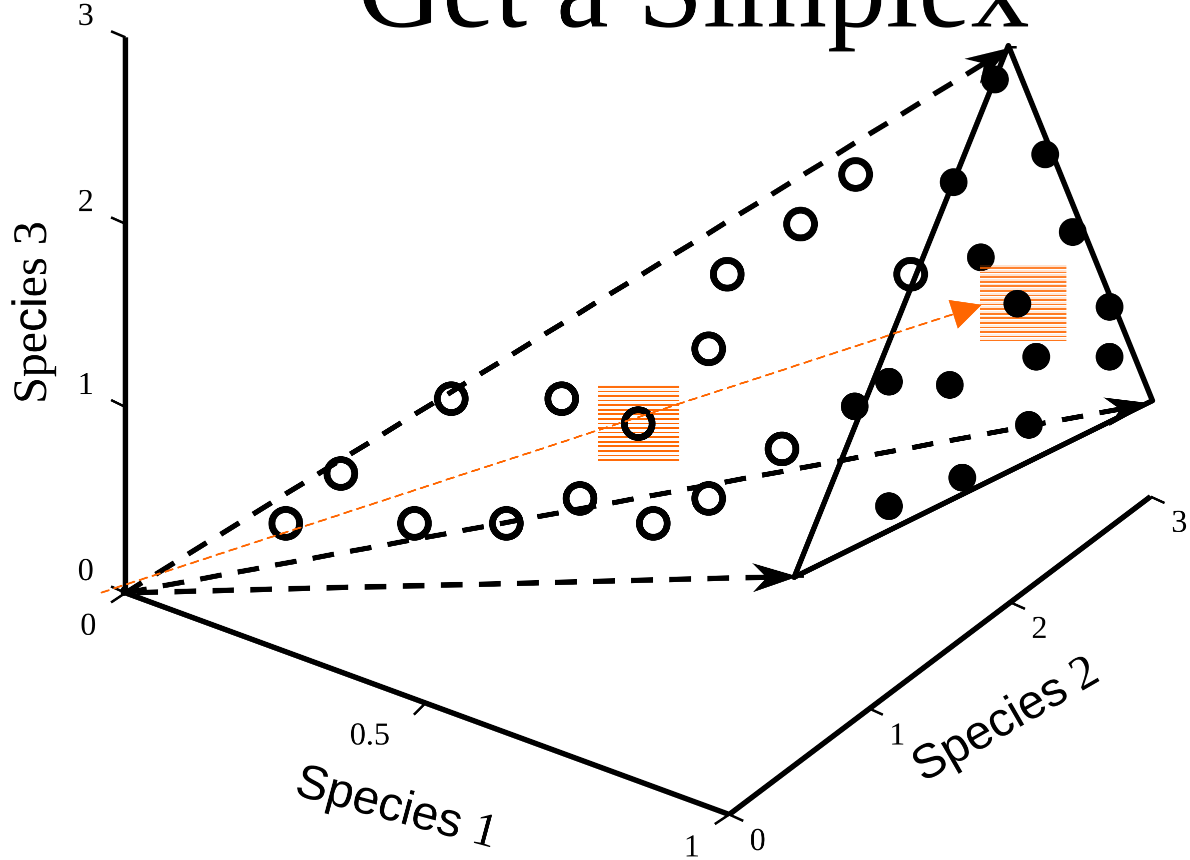
Two Sources - Plane



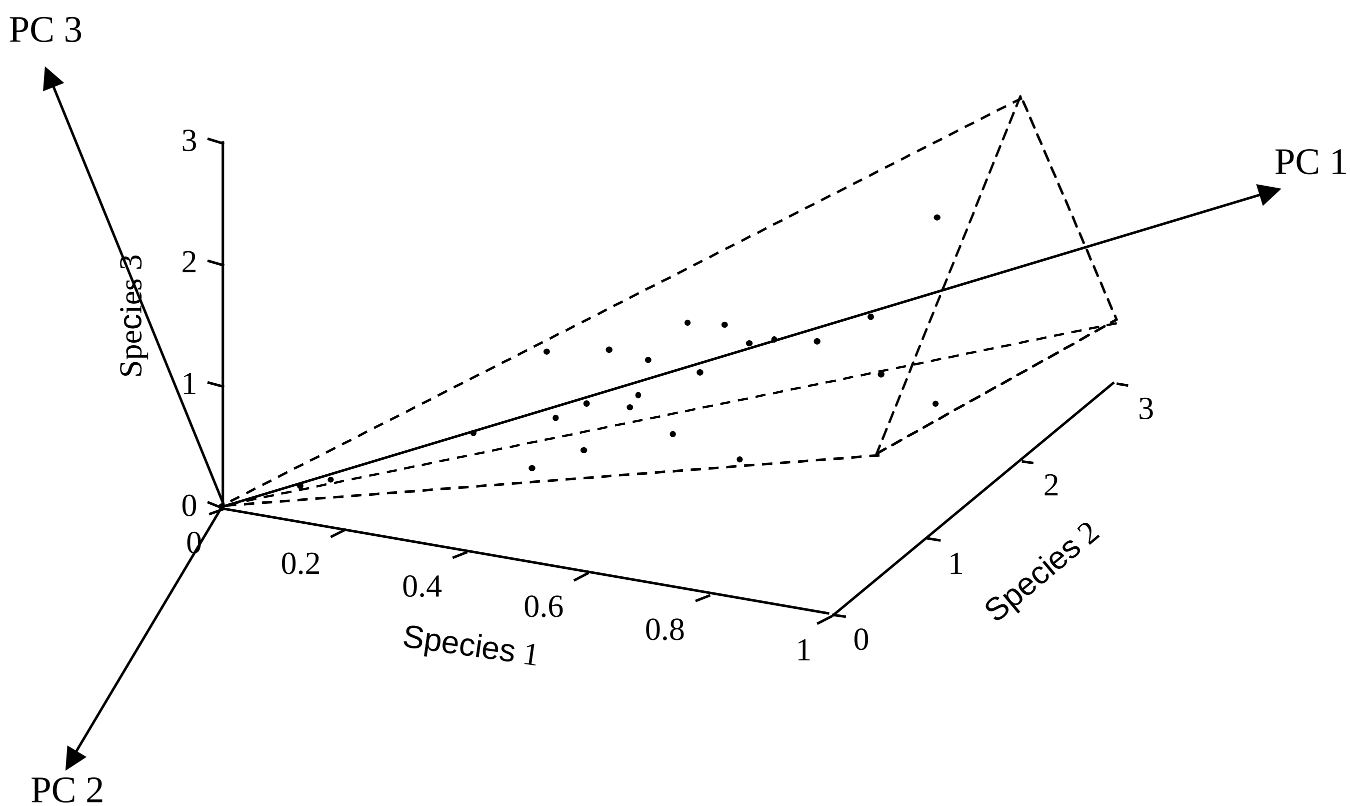
Three Sources



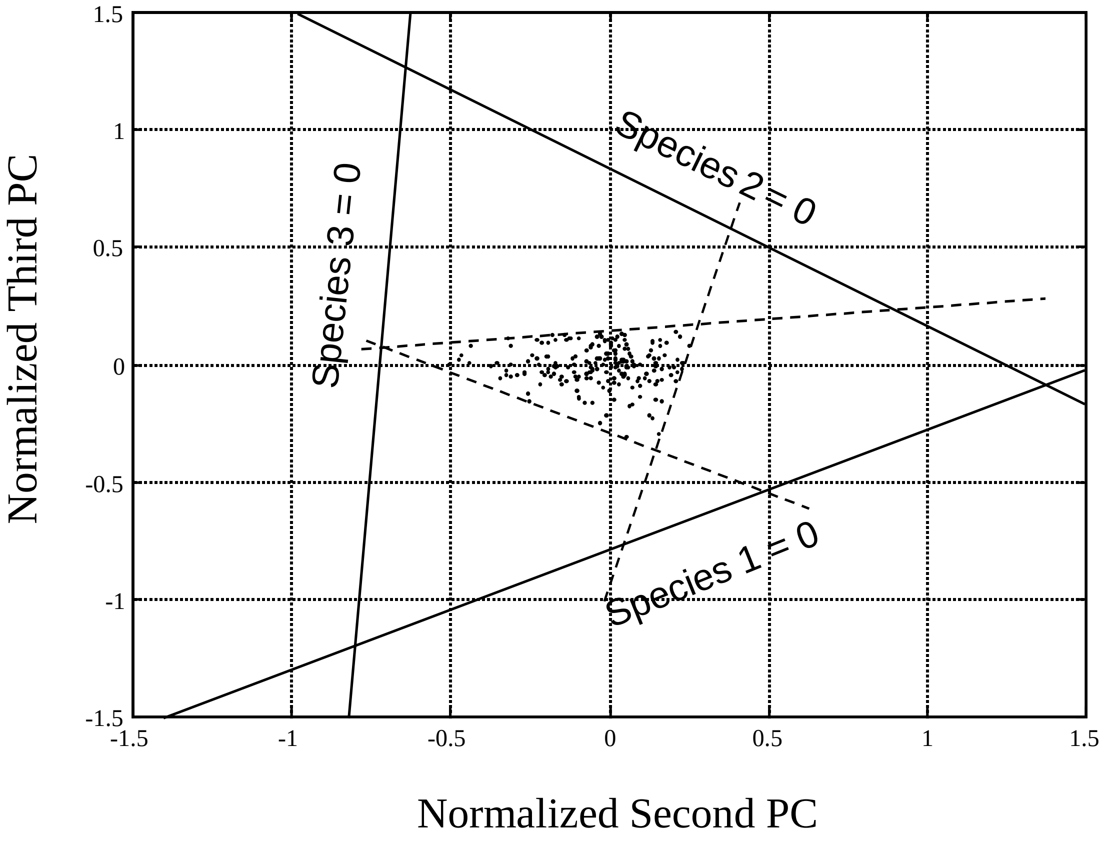
Projection to N-1 Dimensions to Get a Simplex



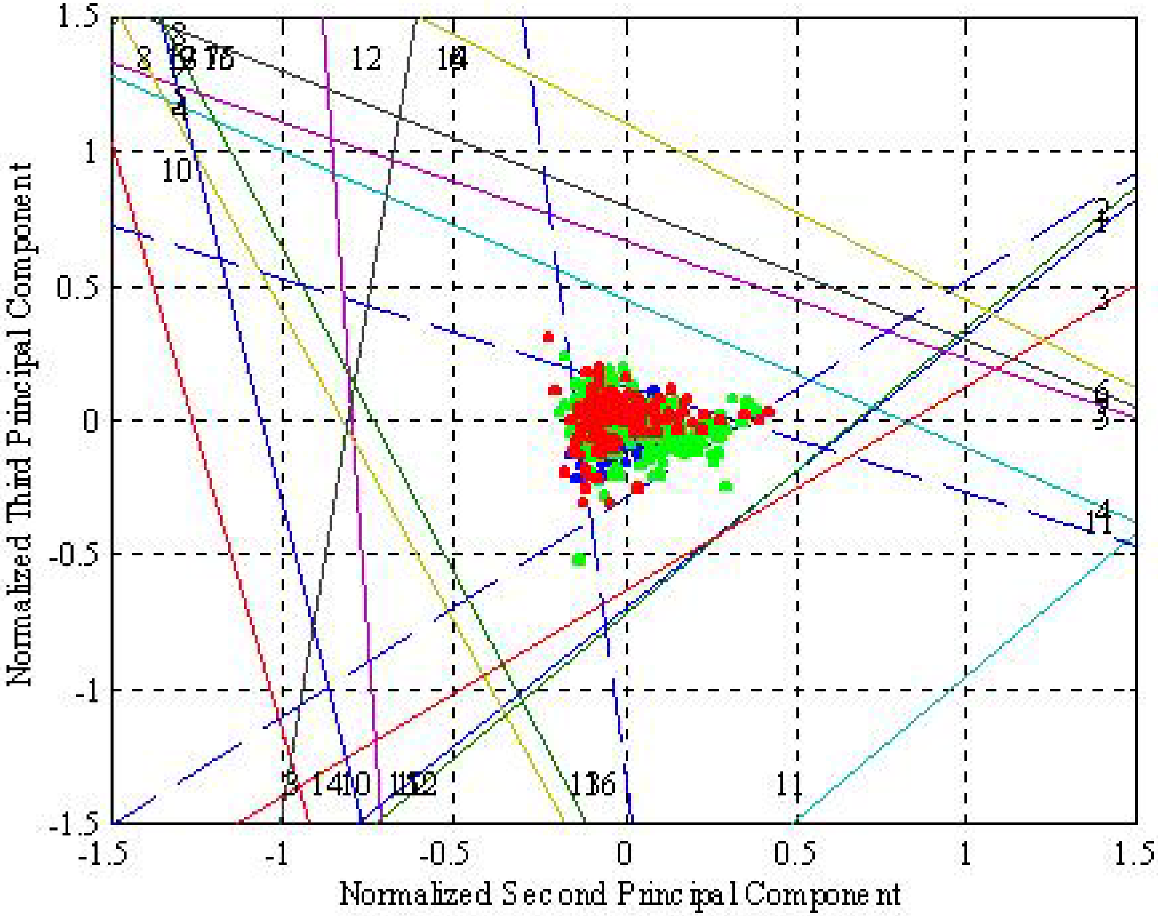
Principal Components



Projection to Plane PC = 1



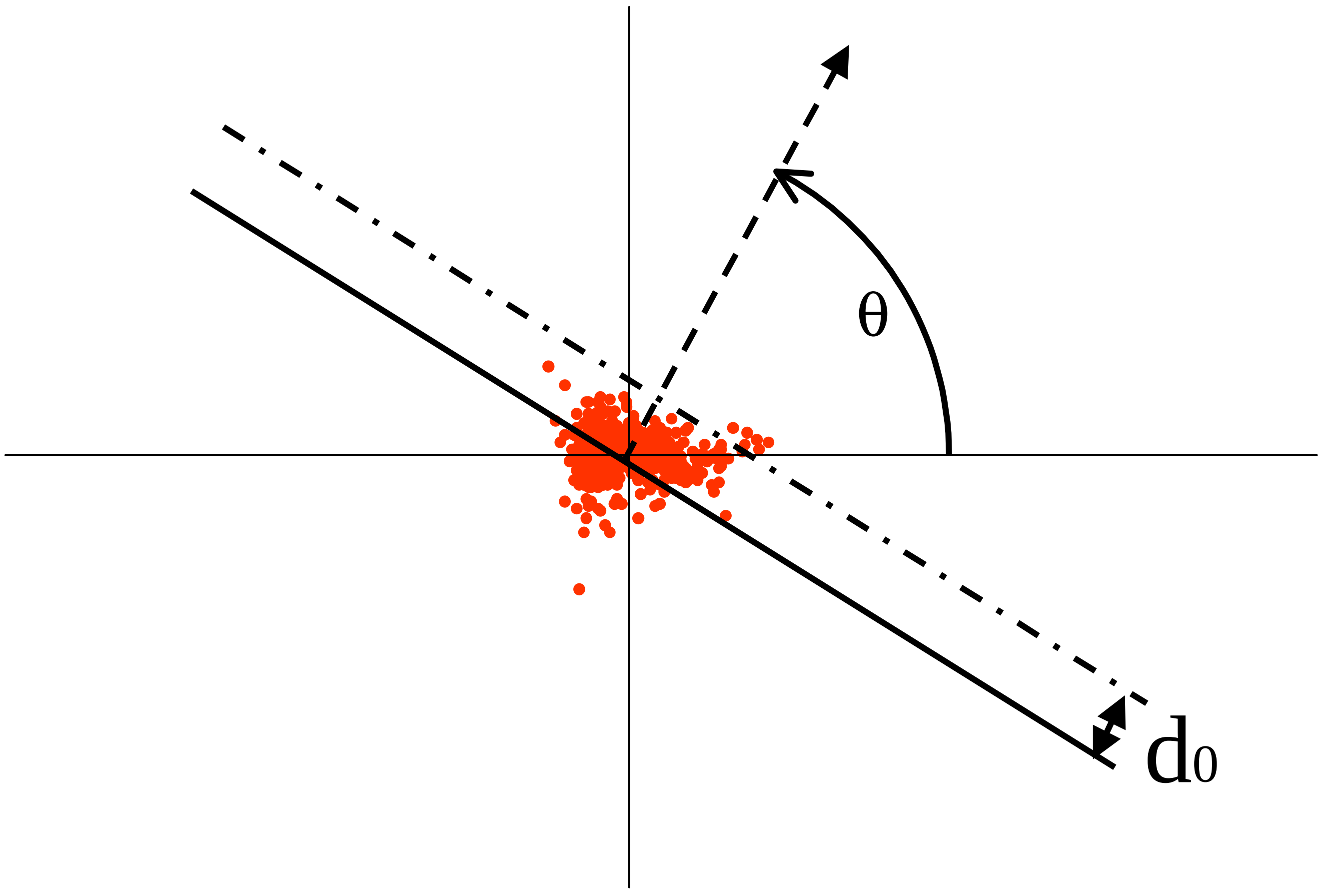
UNMIX 3-D Plot - Atlanta Data



Finding Edges in the Data

More properly, finding hyperplanes
that define a simplex

Parameterizing an Edge



Finding a Subspace Parallel to the Edge

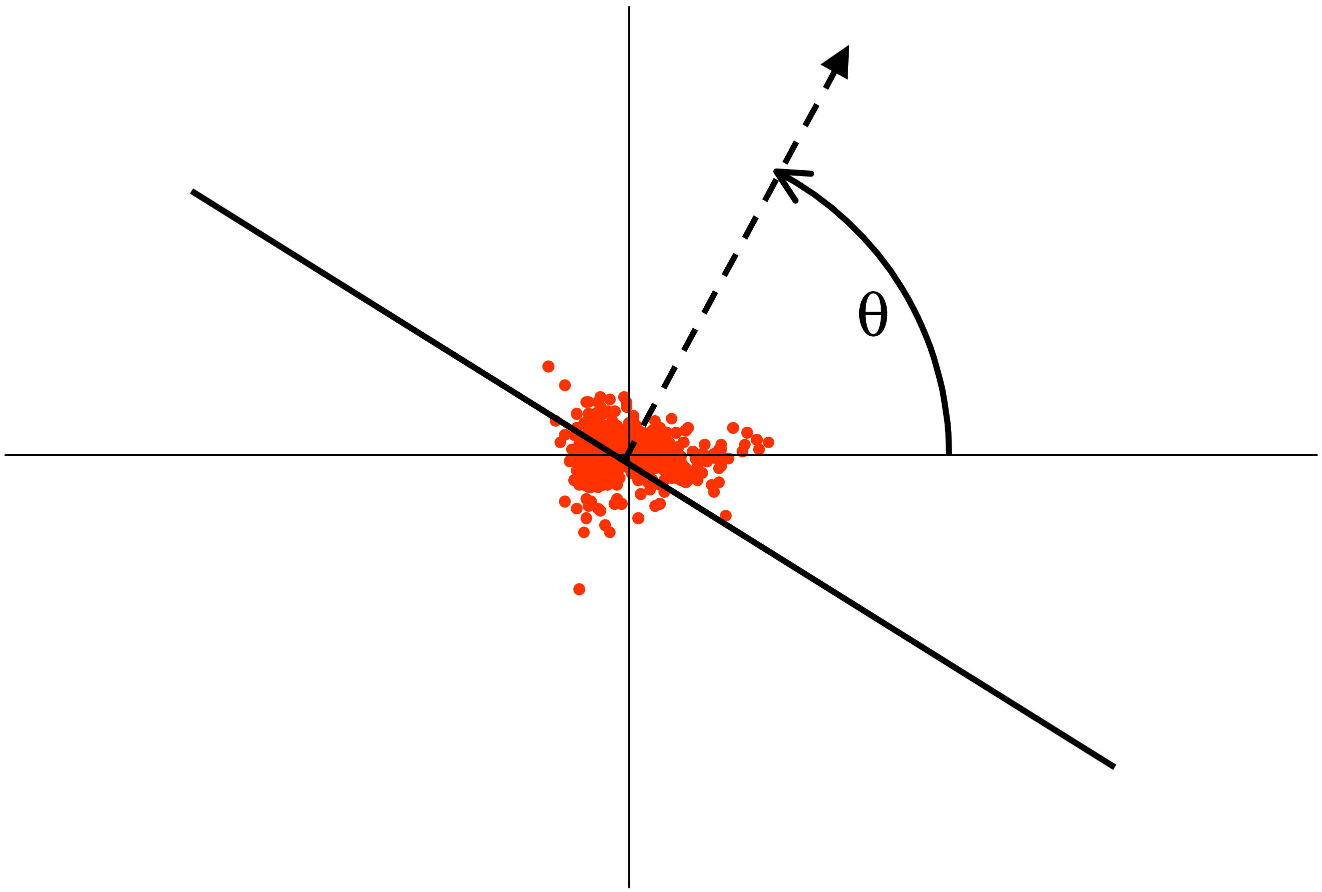
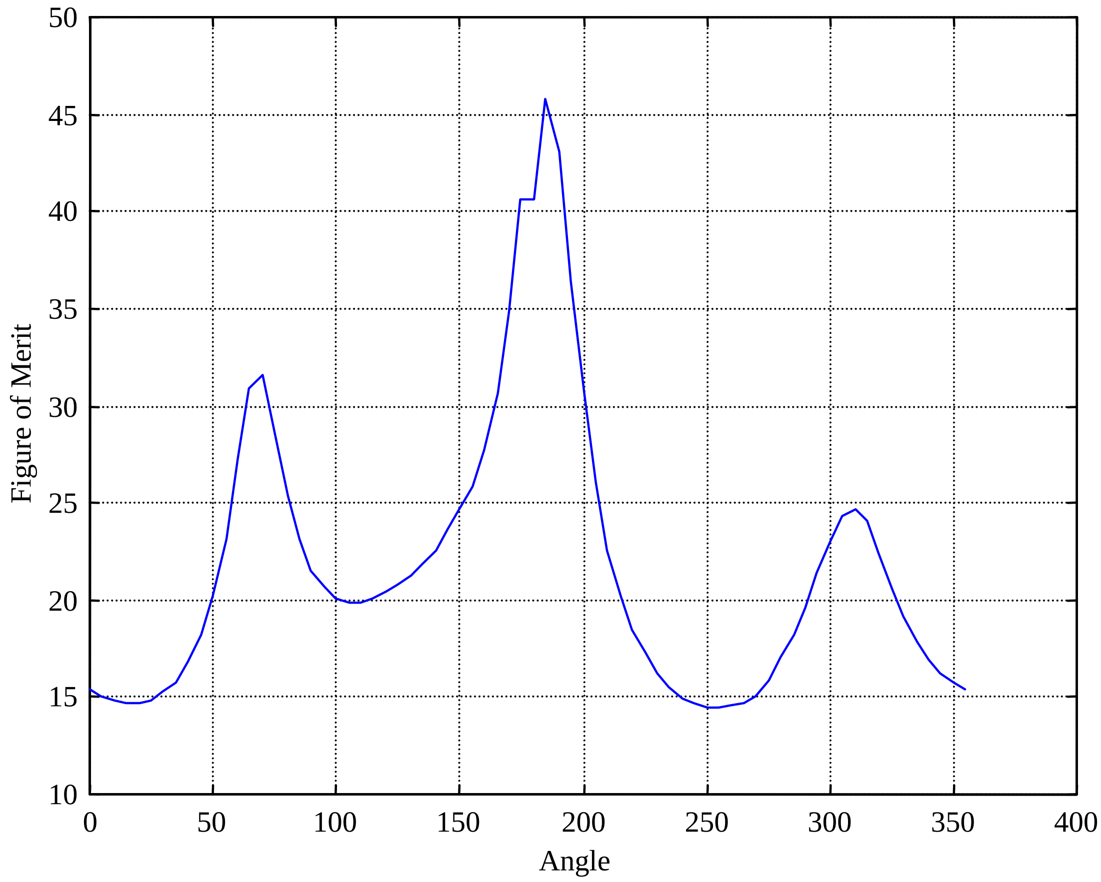


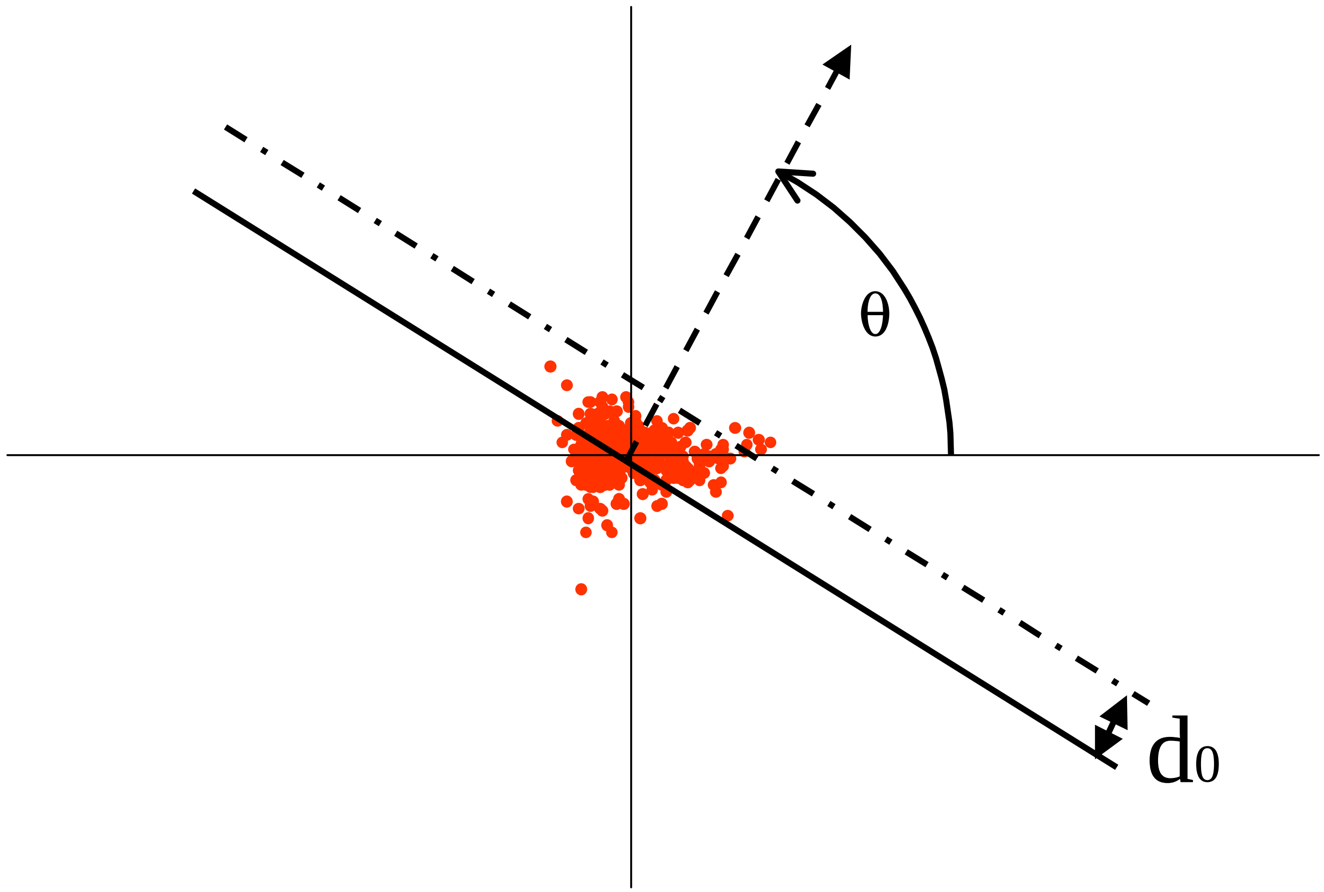
Figure of Merit for Edges

- Find the distance of all points to the given reference line.
- Sort the distances
- Calculate one over the standard deviation of the closest x percent, where x is 5 to 20, but usually 15.

Figure of Merit for Atlanta Data



Parameterizing an Edge



Statistical Model of an Edge

$$D(a,\sigma,d_0) = N(0,\sigma) + U(a) + d_0$$

where

$$D(a,\sigma,d_0) = \text{distance of the point to the edge,}$$

$$N(0,\sigma) = \text{normal distribution, mean 0, std. dev. } \sigma,$$

$$U(a) = \text{uniform distribution on } [0 \text{ a}], \text{ and}$$

$$\mathrm{d}_0 = \text{offset from the origin.}$$

Distribution of Distances from an Edge

Let

$$F(x) = \text{cumlative standard normal distribution}$$

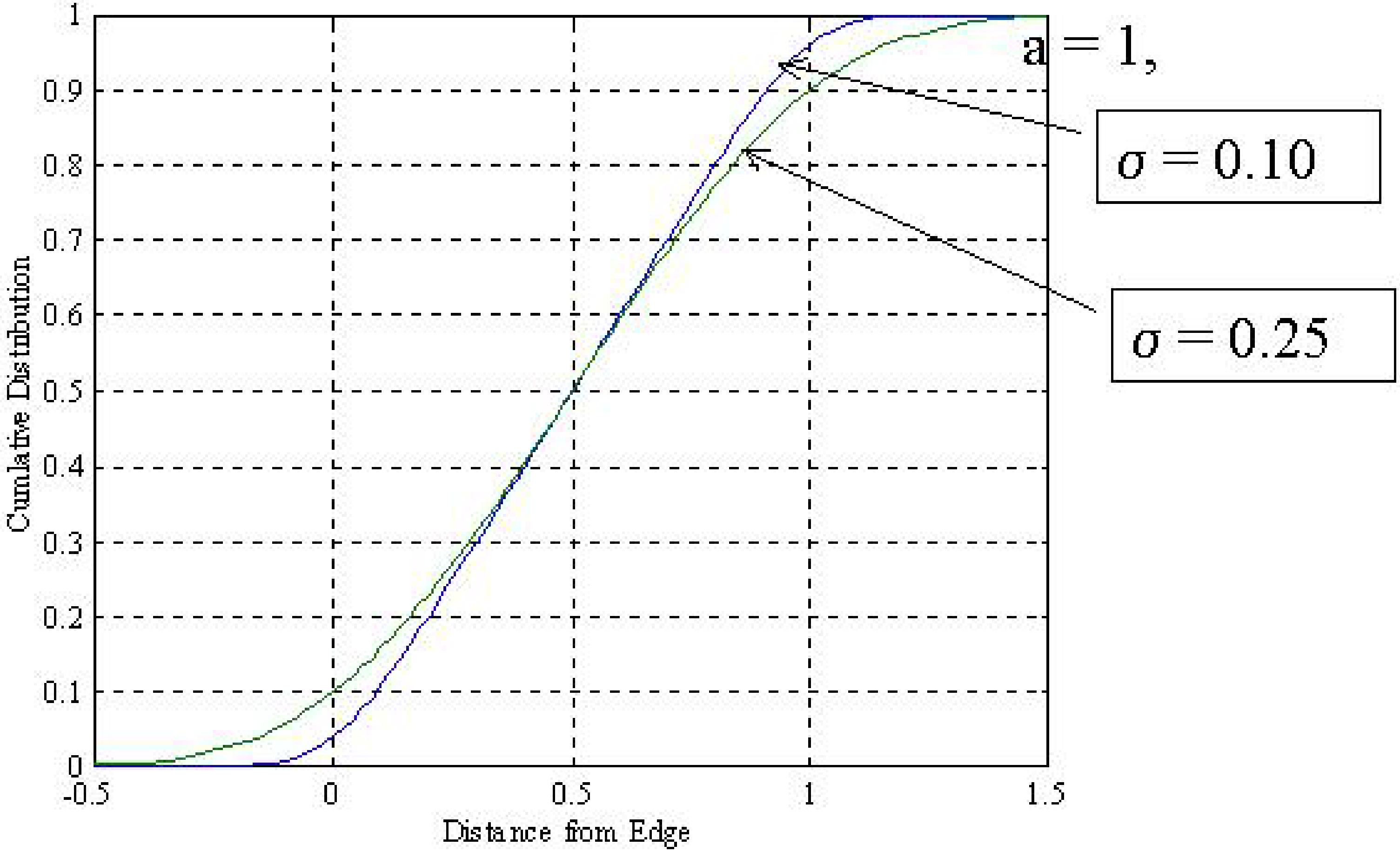
$$= (2\pi)^{-1/2} \int_{-\infty}^{x} \exp(-0.5y^2)\,dy$$

$$\Phi(x) = \int_{-\infty}^{x} F(s)\,ds, \text{ the "iterated cumlative" distribution,}$$

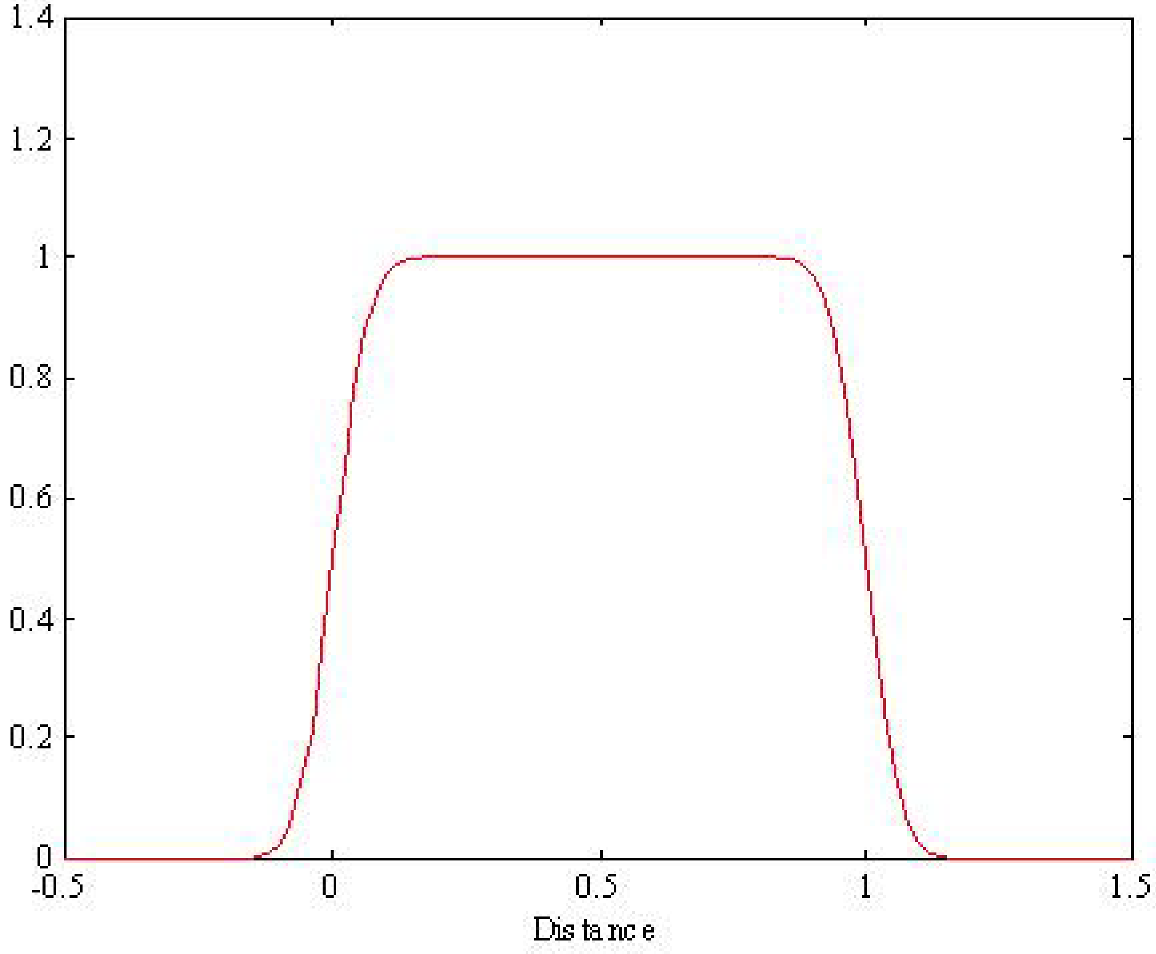
then the cumlative distribution of $\mathrm{D}(\mathrm{a}, \sigma, d_0)$ is

$$G(x, a, \sigma, d_0) = \frac{\sigma}{a}\left[\Phi((x - d_0)/\sigma) - \Phi(((x - d_0) - a)/\sigma\right].$$

Edge Distance Distribution



Edge Distance Density



Assumptions

- Source compositions remain approximately constant
- There are at least N*(N-1) points that have low or no impact from each of the N sources, i.e., need some points with one source missing or low.

Sufficient Conditions for Solution to the Mixture Problem

- If there are n sources, except for error, the data must be confined to a subspace of the data space of dimension equal to n, i.e., the data as a whole is not degenerate.
- The data must contain some observations with each source missing or very low, which define a subspace of dimension n-1.

Advantages

- No assumptions about the number or composition of sources
- No assumptions or knowledge of errors in the data needed
- Automatically corrects source compositions for effects of chemical reactions

Method

- Extension of self-modeling curve resolution to N dimensions (sources)
- Basic idea reference: Henry, R. C. History and Fundamentals of Multivariate Air Quality Receptor Models, 1997. Chemometrics and Intelligent Laboratory Systems. **37**:525-530.

Estimating the Number of Factors by Resampling

- The subspace of data that is spanned by eigenvectors that are not noise dominated does not change much for resampled data
- R.C. Henry, E.S. Park, C.H. Spiegelman, Comparing a new algorithm with the classic methods for estimating the number of factors, *Chemometrics and Intelligent Laboratory Systems* **48**: 91 -97 (1999).

Number of Sources Atlanta Data

	NUMFACT	Eigenvalues of Correlation Matrix	
1	810.9987	15.8856	Rule of 1 gives 1 factor
2	21.9995	0.4922	
3	13.8831	0.3128	Scree test gives 3 factors
4	1.7313	0.0637	Cutoff for NUMFACT is 2.0
5	1.2201	0.06	so it also gives 3 factors
6	1.3044	0.0483	
7	1.1504	0.0353	
8	0.7981	0.0242	
9	0.588	0.0198	Bartlett’s test gives 9 factors
10	0.4458	0.0154	
11	0.3615	0.0125	
12	0.3225	0.0101	
13	0.2652	0.0074	
14	0.1662	0.0049	
15	0.1305	0.0037	
16	0.1056	0.0026	
17	0.0761	0.0015	

UNMIX Model Output

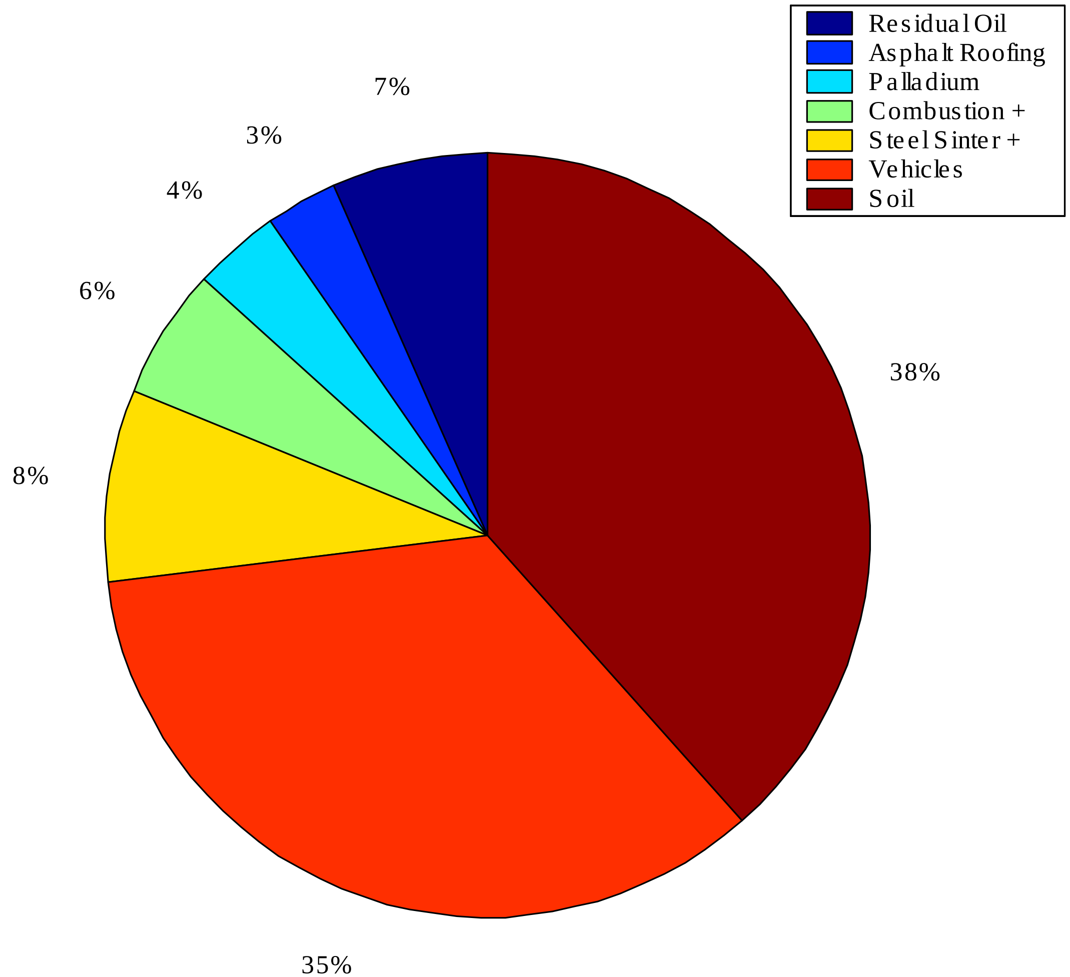
- Number of sources
- Composition of each source
- Source contributions to each sample
- Uncertainties in the source compositions
- Apportionment of the average total mass, if total mass is included in the model.

Simulated Data Results

Sources Other Than Soil and Vehicles

Source	Defining Elements
Asphalt Roofing	Cs, Co
Residual Oil	Ni, V
Combustion	Zn, Br
Steel Sinter +s’blast?	Cu, Cr
Aircraft Jet Fuel	As, NO_3
Unknown	Mg, Pd, Se

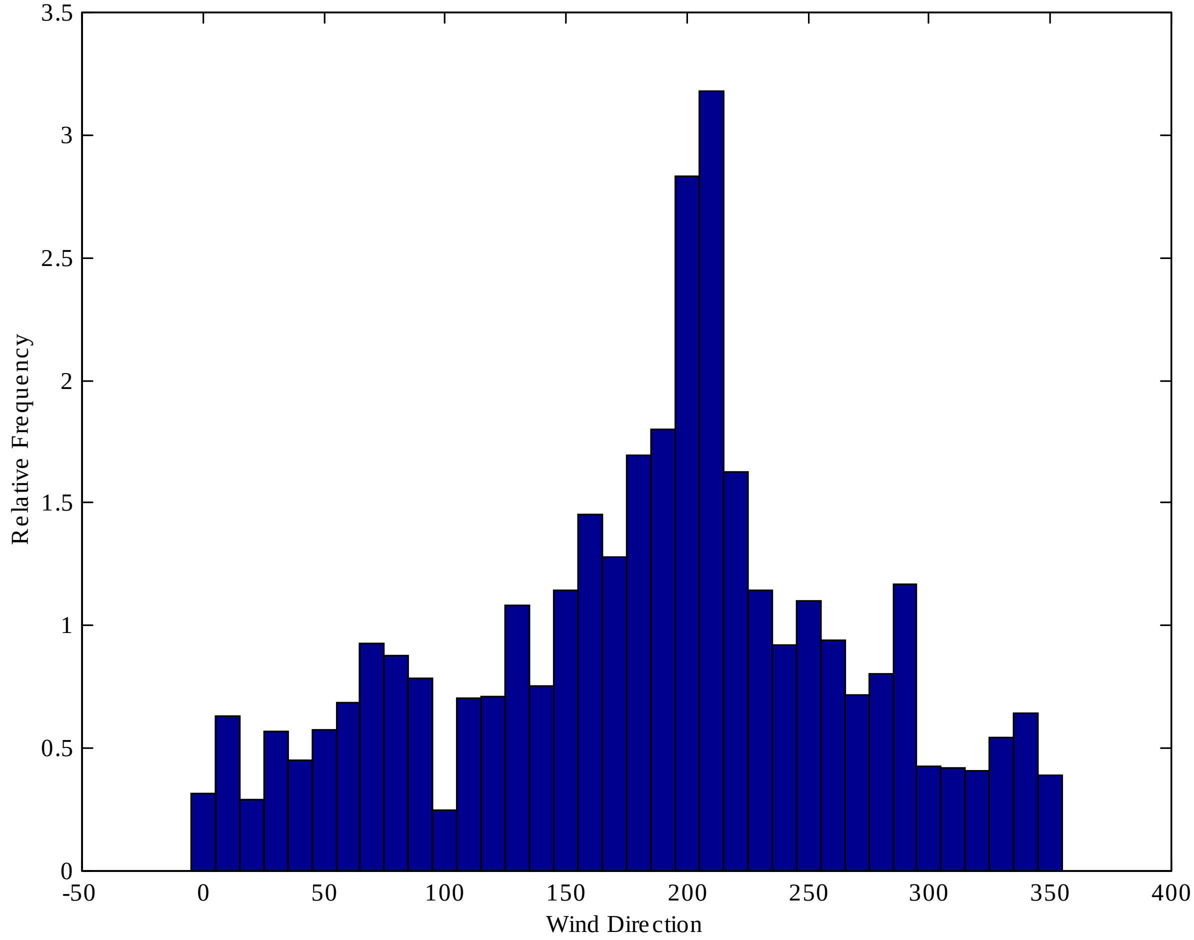
Seven Source Solution



Simulated Data Source Apportionment

	Mean($\mu g/m^3$)	Std. Dev.
Soil	26.9	2.4
Vehicles	24.6	2.3
Residual Oil	6.7	0.8
Combustion	2.8	0.8
Remaining sources	6.5	4.9

Steel Sinter



Direction of Sources

Residual Oil	10 –30
Combustion (broad)	30-50 (60 - 80)
Se (broad)	20 – 40
Steel Sinter +s'blast?	200 –220
Aircraft Jet Fuel	200 –220
Asphalt Roofing	210 – 230
Pd	260 - 280
Mg	215 - 235

Phoenix Data Results

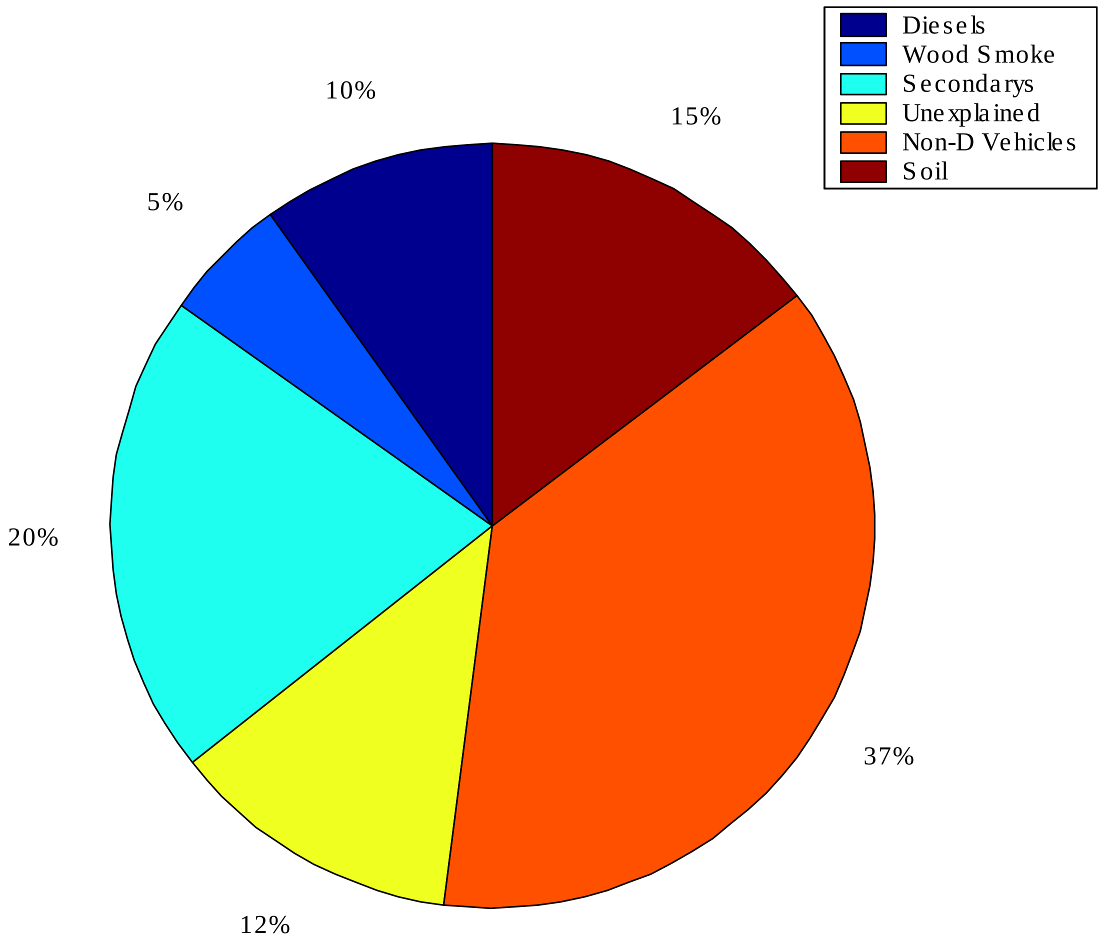
Phoenix Source Compositions

	Diesels	Veg. Burn	Secondary	Unexplained	Vehicles	Soil
PM_FINE	1241	662	2563	1550	4678	1847
AL	0.00057	0.00251	0.00495	0.01139	-0.00089	0.05502
SI	0.01706	0.00637	0.01265	0.03654	-0.00247	0.13751
S	-0.01139	0.00324	0.12599	0.04742	0.00094	0.02573
K	0.00544	0.06400	0.00206	0.00968	0.00112	0.02050
CA	0.01191	-0.00151	0.00392	0.01295	0.00127	0.04749
NON-SOIL K	0.00316	0.06315	0.00037	0.00481	0.00145	0.00217
MN	0.00323	-0.00010	0.00015	0.00033	0.00004	0.00074
FE	0.03832	-0.00460	0.00282	0.01294	0.00871	0.04105
BR	0.00001	0.00031	0.00018	0.00157	0.00016	0.00008
OC	0.27732	0.56208	0.33589	0.48133	0.49149	0.16927
EC	0.30102	0.07751	0.02509	0.05026	0.17192	0.01986

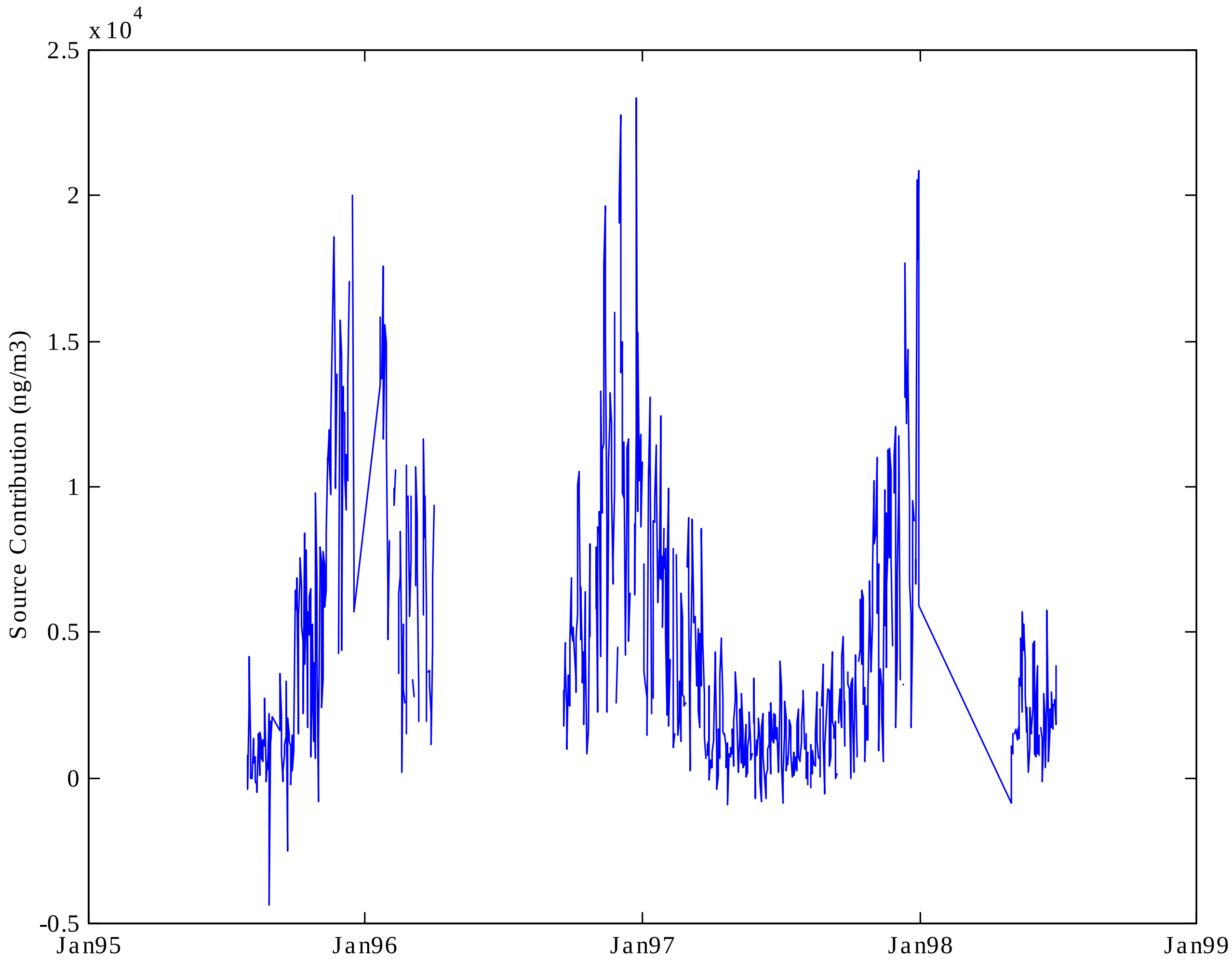
Signal to Noise for Normalized Source Composition

	Diesels	Veg. Burn.	Secondary	Unexplained	Vehicles	Soil
PM_Fine	**4.7**	**2.3**	**11**	**4.9**	**9.7**	**6.7**
AL	0.2	0.2	**5.9**	**4.9**	-1	**6.4**
SI	**2.3**	0.2	**5.9**	**5.8**	-1.2	**6.5**
S	-0.8	0.1	**19.3**	**5.1**	0.4	**4.6**
K	**3.6**	0.4	**5.4**	**6.7**	**2**	**7.5**
CA	**3.9**	-0.1	**4.5**	**6**	1.6	**7**
N-S K	**2.7**	0.4	1.6	**6**	**3.3**	**2.6**
MN	**5**	-0.1	**2.5**	**5.1**	1.1	**6**
FE	**6.3**	-0.1	**2.5**	**7.3**	**12.3**	**7.9**
BR	0	0.9	**9.1**	**6.5**	**9**	1
OC	**5.1**	1.6	**24.3**	**15.4**	**39.1**	**4**
EC	**6.6**	0.4	**2.8**	**2.8**	**23.8**	1.1

Phoenix Source Apportionment



Vehicle Time Series



Secondary Pollutants Time Series